MIDFLORIDA Credit Union Capstone Project: Branch Site Selection

Problem Statement and Objectives

- Harness public demographic data, traffic metrics, competition analysis, and insights from existing branches to design a model pinpointing the most strategic locations for new branches
- Goal: Explain the model's criteria and showcasing at least one site recommendation for potential expansion
- Determine what makes a site profitable and what locations have those characteristics

Customer Needs

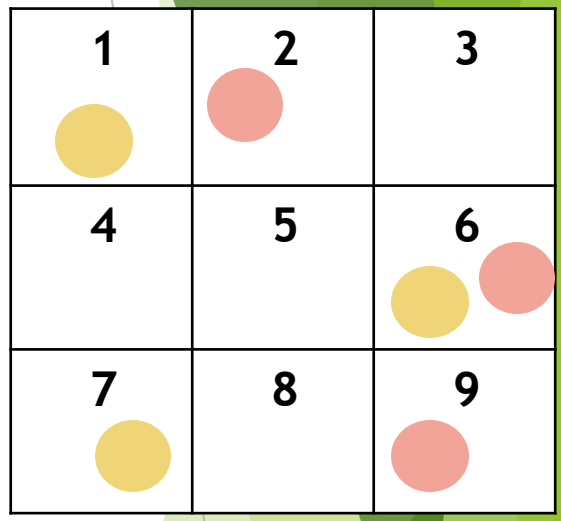
- Primary
  - Create models that can be cross-referenced to identify potential branch sites for MidFlorida
  - Identification of variables that affect profitability and the success of a branch
- Secondary
  - The model is simple enough to explain to higher management
  - Finding data that is useful and available for determining profitability
- Latent
  - Balancing a model that can be as accurate as possible while not overfitting

Solution Development

- MidFlorida's current approach: comparing the features of potential branch sites to the features of their current locations
  - Statistically cannot produce usable results due to sample selection
- Our approach: use game theory as a framework for:
  - Market saturation of financial institutions for each census tract
  - MidFlorida's strategy
- Heat maps and qualitative figures used to ensure interpretability and ease of understanding

Game Theory

- In game theory, we have strategic rational players that are attempting to optimize their outcome (profit) based on the decisions they expect other players to make optimally
- Let g represent census tracts in Florida (the area of interest):
  - $g = \{1, 2, \ldots, G\}$
- Let i represent different financial institutions:
  - $i = \{1, 2, \ldots, N\}$
- Each institution i has a strategy:
  - $b_i = (0, 1, 1, 1, \ldots, 0)$
- Where:
  - $b_{ig} \begin{cases} 1 & \text{if a branch for institution } i \text{ is open at location } g \\ 0 & \text{else} \end{cases}$



g = 1,2,3,4,5,6,7,8,9

b_{yellow} = {1,0,0,0,0,1,1,0,0}

b_{red} = {0,1,0,0,0,1,0,0,1}

i = {yellow, red}

Figure 1: A simple example of strategy

Game Theory

- The optimal (Nash Equilibrium) strategy for all competitors can be written as matrix B*:
- To determine potential profitability, estimate:
  - Sum of each column g:

$$n_g = \sum_{i=1}^{N} \mathrm{b_{ig}} = \text{Number of institutions present in census tract } g$$

  - The strategy for MidFlorida: $b_{(MidFlorida)g}$

$$n_i = \sum_{g=1}^{G} b_{(MidFlorida)g} = \text{Number of MidFlorida branches}$$

  - The profit of an institution i is dependent on their strategy, b_i, the strategy of all other institutions, b_{-i}, and a matrix X of relevant predictors and the location of interest

$$\pi_i(b_i, b_{-i}); X$$

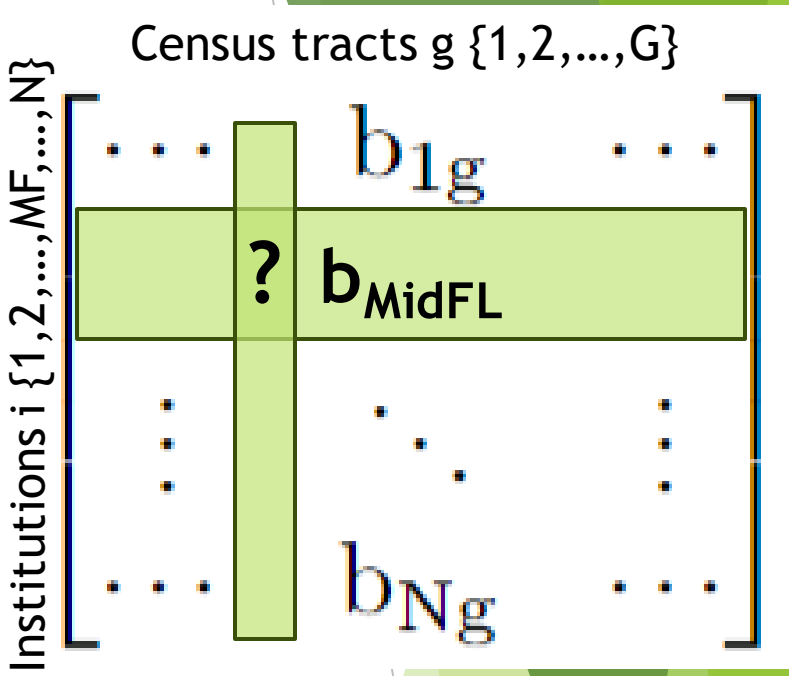


Figure 2: Matrix B*, a combination of locations and institutions

Heckman Theory

- **Selection:**

$$Z_i^* = \gamma X_i + \delta W_i + u_i$$

$$Z_i = \begin{cases} 1 \; if \; Z_i^* > 0 \\ 0 \; if \; Z_i^* \leq 0 \end{cases}$$

Where:

Z_i^* *is an unobserved variable that determines propensity for observation*

Z_i *is the binary variable that represents selection (Accepted tracts)*

γ *and* δ *are estimated parameter vectors*

X_i *is a vector of explanatory variables*

W_i *is a vector of insrtamental variables*

u_i *is the error term*

Stage 1: Predicts whether location i will be selected/observed

- **Outcome:**

$$Y_i = \beta X_i + \delta \lambda \left(\frac{CDF}{PDF}\right) + v_i$$

Where:

Y_i *is the predicted dependent variable (# of checking accounts)*

β *is an estimated parameter vector*

X_i *is a vector of explanatory variables*

δ *is the coefficient of the inverse Mill's ratio*

$\lambda\left(\frac{CDF}{PDF}\right)$ *is the inverse Mill's ratio (a function of the cumulative density function and the probability density function)*

v_i *is the error term*

Stage 2: Predicts the number of checking account a branch will have in census tract i conditional on whether or not that location was selected

Heckman Results

- Selection:

	Model 1	Model 2	Model 3	Model 4	Model 5
lnPopulation_Density	0.037	0.037	0.037	0.064	0.037
	(0.057)	(0.057)	(0.057)	(0.052)	(0.057)
lnMedian_Home_Value	-0.043	-0.043	-0.043		-0.043
	(0.158)	(0.158)	(0.158)		(0.158)
lnFair_Share_Pop	-0.852	-0.852	-0.852	-0.755	-0.852
	(0.109)	(0.109)	(0.109)	(0.099)	(0.109)
lnemployment_ratio	0.497	0.497	0.497	0.368	0.497
	(0.217)	(0.217)	(0.217)	(0.191)	(0.217)
lnMedian_Household_Income	0.162	0.162	0.162	0.157	0.162
	(0.229)	(0.229)	(0.229)	(0.155)	(0.229)
cu_ratio	2.202	2.202	2.202	2.076	2.202
	(0.215)	(0.215)	(0.215)	(0.196)	(0.215)
Partial_5mi	-0.071	-0.071	-0.071	-0.060	-0.071
	(0.255)	(0.255)	(0.255)	(0.250)	(0.255)
Intercept	1.115	1.115	1.115	-0.096	1.115
	(1.976)	(1.976)	(1.976)	(1.870)	(1.976)
lambda	-0.219	-0.382	-0.214	-0.142	-0.127
	(0.262)	(0.271)	(0.259)	(0.271)	(0.268)

- Outcome:

	Model 1	Model 2	Model 3	Model 4	Model 5
lnPopulation_Density	-0.117	-0.227	-0.106		-0.052
	(0.123)	(0.124)	(0.122)		(0.124)
lnMedian_Home_Value	-0.603		-0.882	-0.916	-0.551
	(0.215)		(0.288)	(0.281)	(0.221)
lnRenter_Ratio	-0.303	-0.152		-0.160	-0.184
	(0.175)	(0.177)		(0.213)	(0.174)
lnFair_Share_Pop	-0.191	-0.122	-0.127	-0.197	-0.106
	(0.209)	(0.220)	(0.191)	(0.203)	(0.213)
lnemployment_ratio	-0.836	-0.731	-0.638	-0.646	
	(0.362)	(0.384)	(0.345)	(0.356)	
lnMedian_Household_Income			0.779	0.591	
			(0.404)	(0.491)	
Intercept	18.174	11.587	13.007	14.717	15.842

	Model 1	Model 2	Model 3	Model 4	Model 5
AIC	646.74	652.04	647.66	658.88	649.86
BIC	750.85	749.65	745.27	756.79	747.47

Figure 3: Table of Heckman Results along with AIC and BIC

Heckman Results

We can see that selection is a big deal by the difference in predictions where MidFlorida's are located vs. where they aren't → AT LEAST need to be confident that the Heckman model works or find a new approach

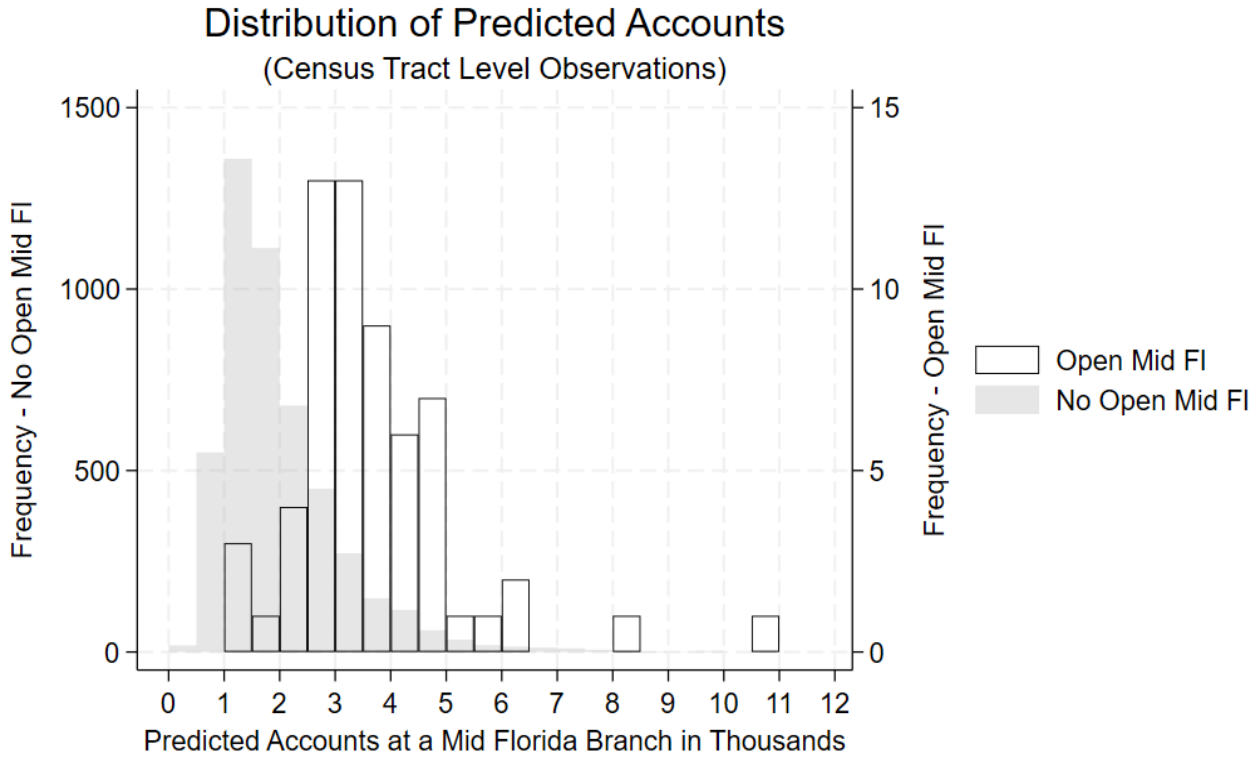


Figure 4: The Distribution of Predicted Checking Accounts

Selection:

$$Z_i = \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 X_4 + \gamma_5 X_5 + \gamma_6 X_6 + \gamma_7 X_7 + \delta W_1 + \delta W_2 + u_i$$

$\gamma_1 = 0.037$ $X_1 = Pop\ Density$

$\gamma_2 = -0.043$ $X_2 = Median\ Home\ Value$

$\gamma_3 = -0.852$ $X_3 = Fair\ Share\ Pop$

$\gamma_4 = 0.497$ $X_4 = Employment\ Ratio$

$\gamma_5 = 0.162$ $X_5 = Median\ Household\ Income$

$\boldsymbol{W_1 = -0.071}$ $\boldsymbol{X_2 = Partial\ 5mi}$

$\boldsymbol{W_2 = 2.202}$ $\boldsymbol{X_2 = Credit\ Union\ Ratio}$

Outcome:

$$Y_i = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + X_5\beta_5 + X_5\beta_5 + X_6\beta_6 + X_7\beta_7 + \delta\lambda\left(\frac{PDF}{CDF}\right) + v_i$$

$\beta_1 = -0.106$ $X_1 = Pop\ Density$

$\beta_2 = -0.882$ $X_2 = Median\ Home\ Value$

$\beta_3 = -0.127$ $X_3 = Fair\ Share\ Pop$

$\beta_4 = -0.638$ $X_4 = Employment\ Ratio$

$\beta_5 = 0.779$ $X_5 = Median\ Household\ Income$

$\delta = -0.214$ $\lambda\left(\frac{PDF}{CDF}\right) = Inverse\ Mill's\ Ratio$

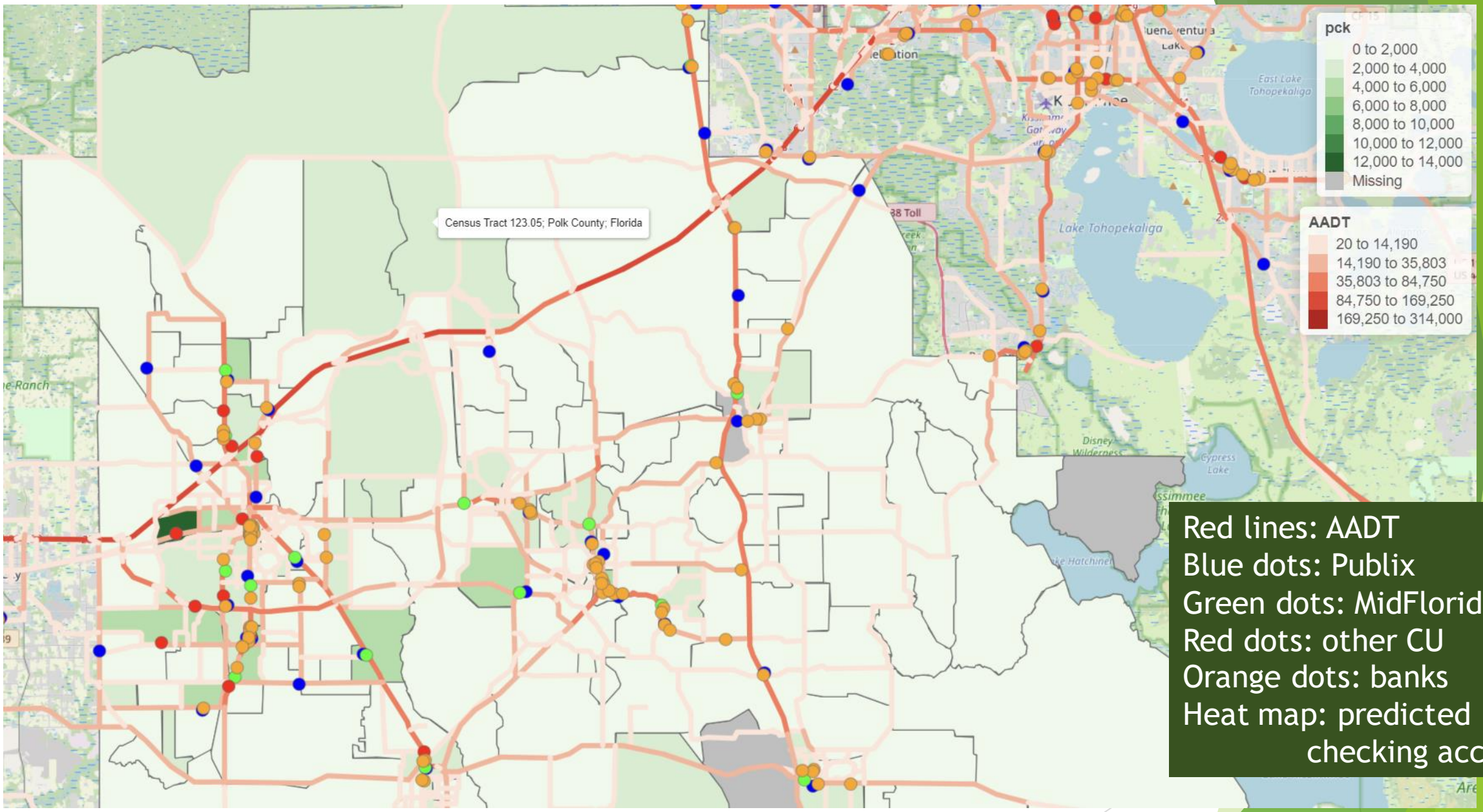


Figure 5: Map illustrating the number of predicted checking accounts by census tract, Polk County

Estimating Number of Financial Institutions per Census Tract

Market Saturation Model

- **Determining demographic indicators to include**
- V-select on multiple linear regression model predicting break even number of financial institutions per capita:

$$Y = \frac{\text{Number of Financial Institutions} + 1}{\text{Total Population}}$$

Model	# Predictors	AIC	BIC	RMSE
1	5	8762.568	8801.532	0.59314
2	7	8749.646	8801.597	0.59113
3	11	8756.216	8842.637	0.59219

Figure 6: Model Selection Criteria

Market Saturation Model

- Poisson regression model

$$\log(\widehat{Y}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7$$

$\beta_0 = -5.884,\ \text{SE} = 0.716$	$Y =$ Number of Financial Institutions
$\beta_1 = 0.492,\ \text{SE} = 0.034$	$X_1 =$ Log Population Density
$\beta_2 = 0.611,\ \text{SE} = 0.033$	$X_2 =$ Log Square Miles
$\beta_3 = 0.560,\ \text{SE} = 0.035$	$X_3 =$ Log Bachelor's Degree Population Ratio
$\beta_4 = 0.547,\ \text{SE} = 0.034$	$X_4 =$ Log Elderly Population Ratio
$\beta_5 = 0.779,\ \text{SE} = 0.032$	$X_5 =$ Log Renter Occupied Housing Ratio
$\beta_6 = 0.402,\ \text{SE} = 0.056$	$X_6 =$ Log Median Household Income
$\beta_7 = -0.095,\ \text{SE} = 0.031$	$X_7 =$ Log Owner Occupied with a Mortgage Housing Ratio

Market Saturation Model

Indicators of potential profitability

- **Probabilities using Poisson distribution**
  - Probability that the number of financial institutions at equilibrium is greater than what currently exists

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

X = Number of Financial Institutions at Equilibrium

k = Actual Number of Financial Institutions

λ = Expected (Fitted) Value

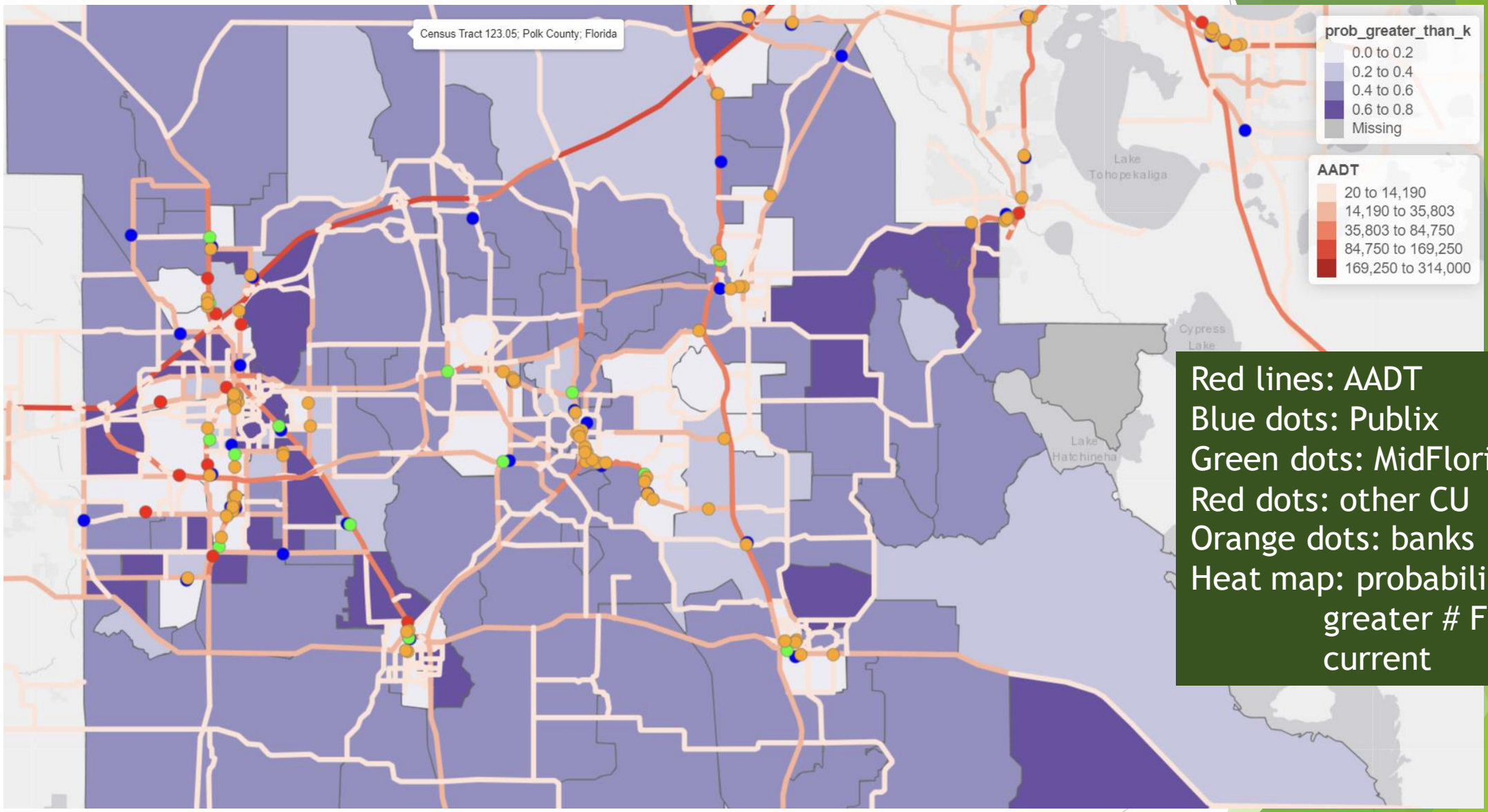


Figure 7: Heat map indicating probability of greater number of financial institutions than exist by census tract, Polk County

Predicted Checking Accounts and Probability Map of Polk County

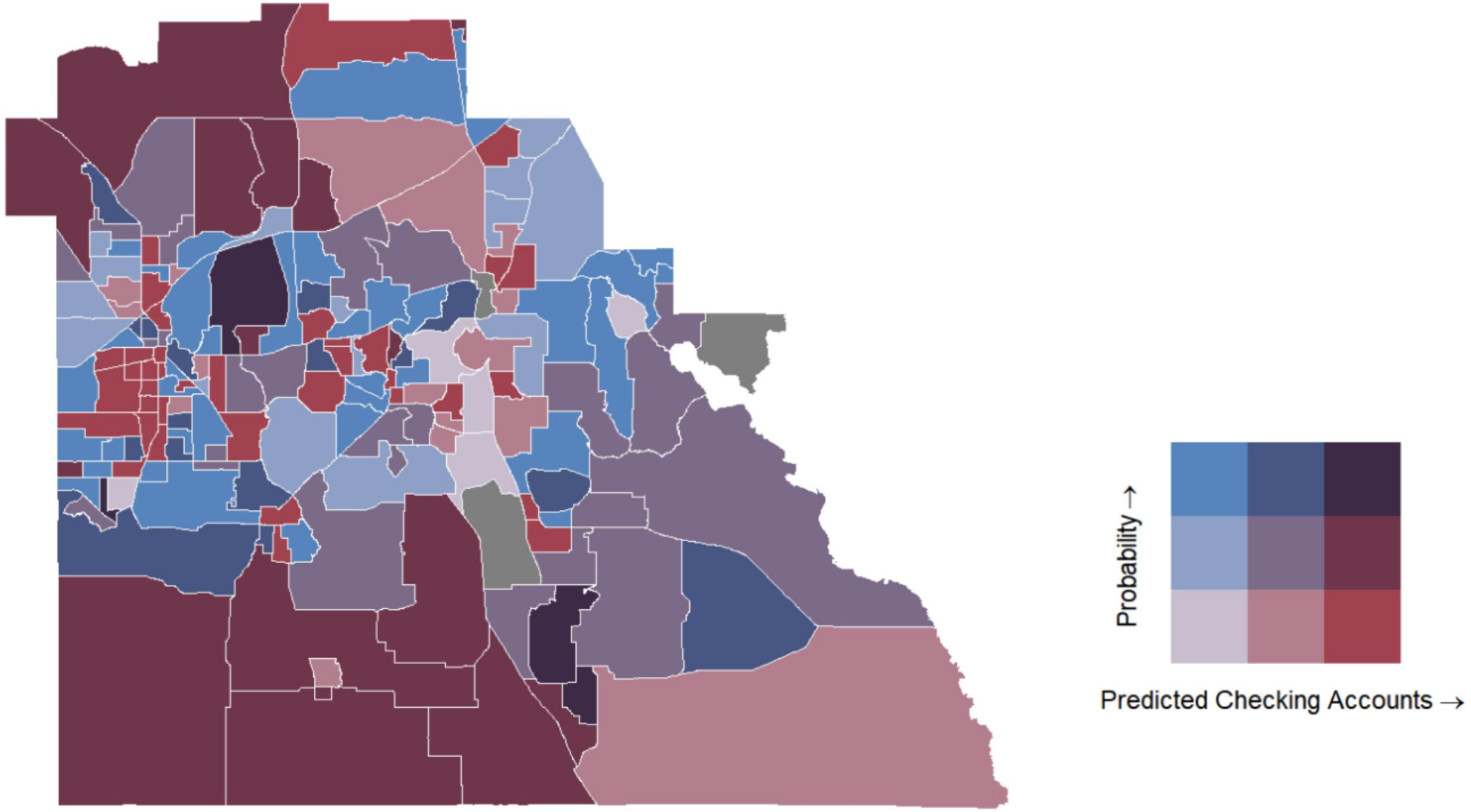


Figure 8: Polk county heat map indicating probability of greater number of financial institutions than exist and predicted number of checking accounts by census tract

Estimating MidFlorida's Strategy: The Most Profitable Locations

Frequent Locations

Automated - Businesses by All Banks (Successful MidFlorida, Wells Fargo, Suncoast, Truist)			
Dollar General	Discount store	22 out of 30	73.3%
Dollar Tree	Discount store	20 out of 30	66.7%
Winn-Dixie	Grocery store	13 out of 30	43.3%
Manual - Businesses by All Banks (Successful MidFlorida, Wells Fargo, Suncoast, Truist)			
McDonald's	Fast food restaurant	30 out of 30	100.0%
Publix	Grocery store	26 out of 30	86.7%
CVS	Convenience store	25 out of 30	83.3%
Advance Auto Parts	Auto parts store	24 out of 30	80.0%
Automated - Businesses by Successful but Not by Less Successful (MidFlorida)			
Beef 'O' Brady's	Family restaurant	7 out of 15	46.7%
Save A Lot	Grocery store	6 out of 15	40.0%
Burger King	Fast food restaurant	5 out of 15	33.3%
Manual - Businesses by Successful but Not by Less Successful (MidFlorida)			
Cheddar's Scratch Kitchen	Family restaurant	8 out of 15	53.3%
Tractor Supply Co	Home improvement store	5 out of 15	33.3%
Sonny's BBQ	Family restaurant	4 out of 15	26.7%

Figure 9: Summary of Web Scraping Results for Automated and Manual Method

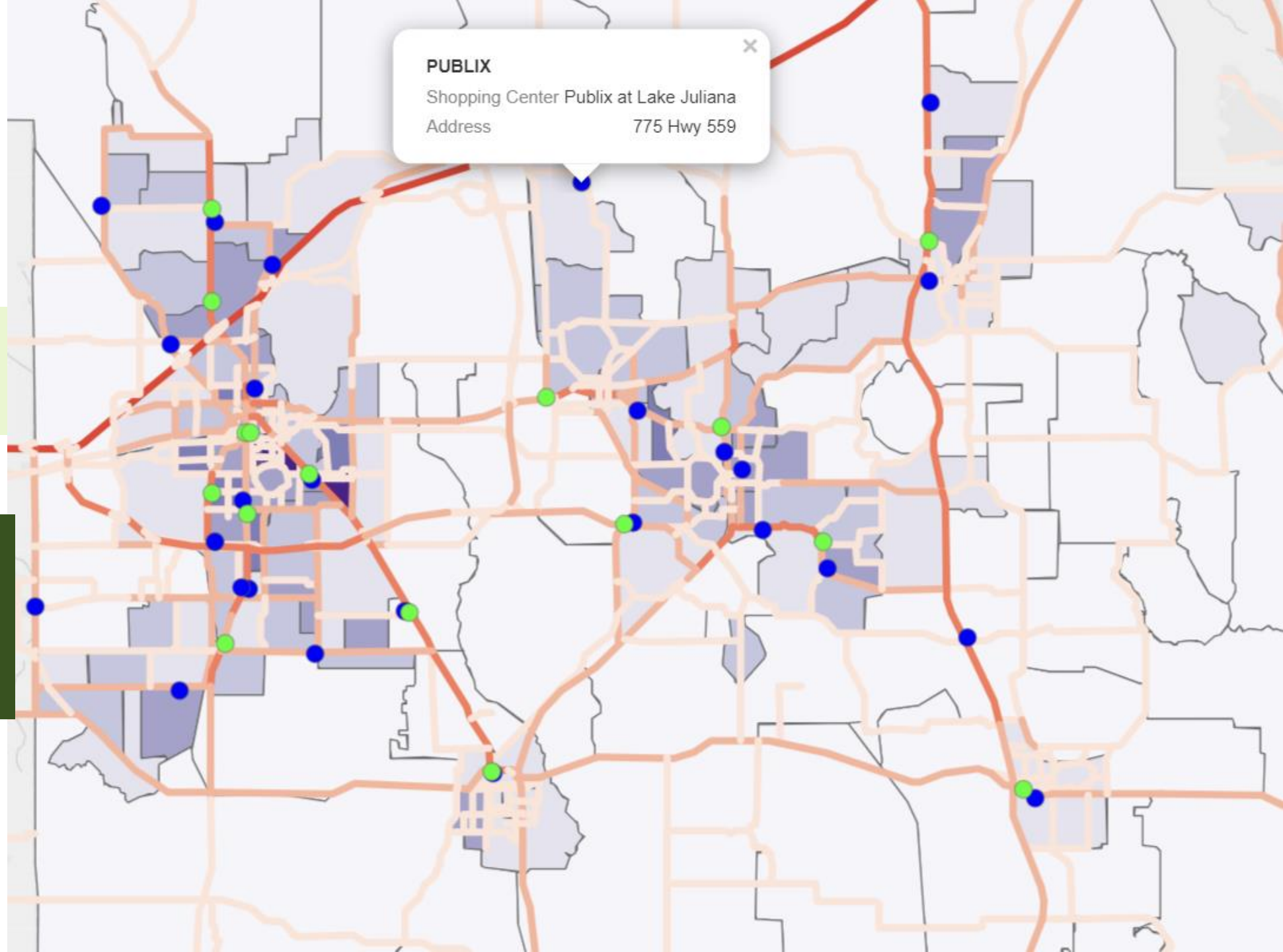
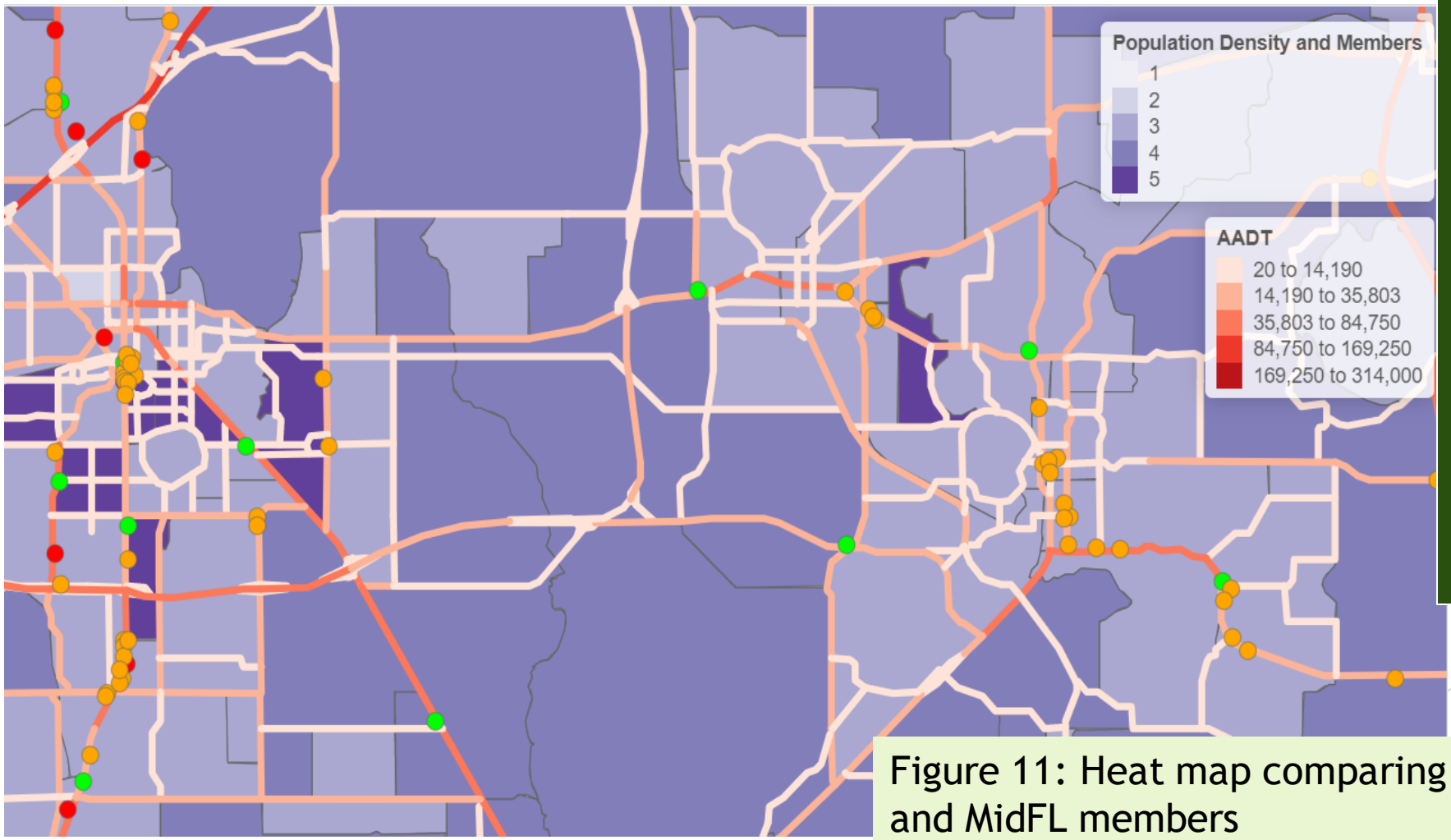


Figure 10: Census Tract with Publix and MidFlorida Locations

Red lines: AADT
Blue dots: Publix
Green dots: MidFlorida
Heat map: population density

Spatial Competitive Threats and Opportunity Analysis



Red lines: ADDT
Green dots: MidFlorida
Red dots: Other Credit Unions
Orange dots: Banks
Heat Map: Pop. Density and Members
1: low population density and low members
2: low population density and high members
3: both are not too low or too high
4: high population density and low members
5: high population density and high members

Figure 11: Heat map comparing population density and MidFL members

POC: Selected Location

- Located at an intersection in the Orange County Area
- High probability of being underserved
- High Predicted Checking Accounts
- High AADT (annual average daily traffic)
- Publix location nearby

Red lines: AADT
Blue dots: Publix
Green dots: MidFlorida
Red dots: other CU
Orange dots: banks
Heat map: population density

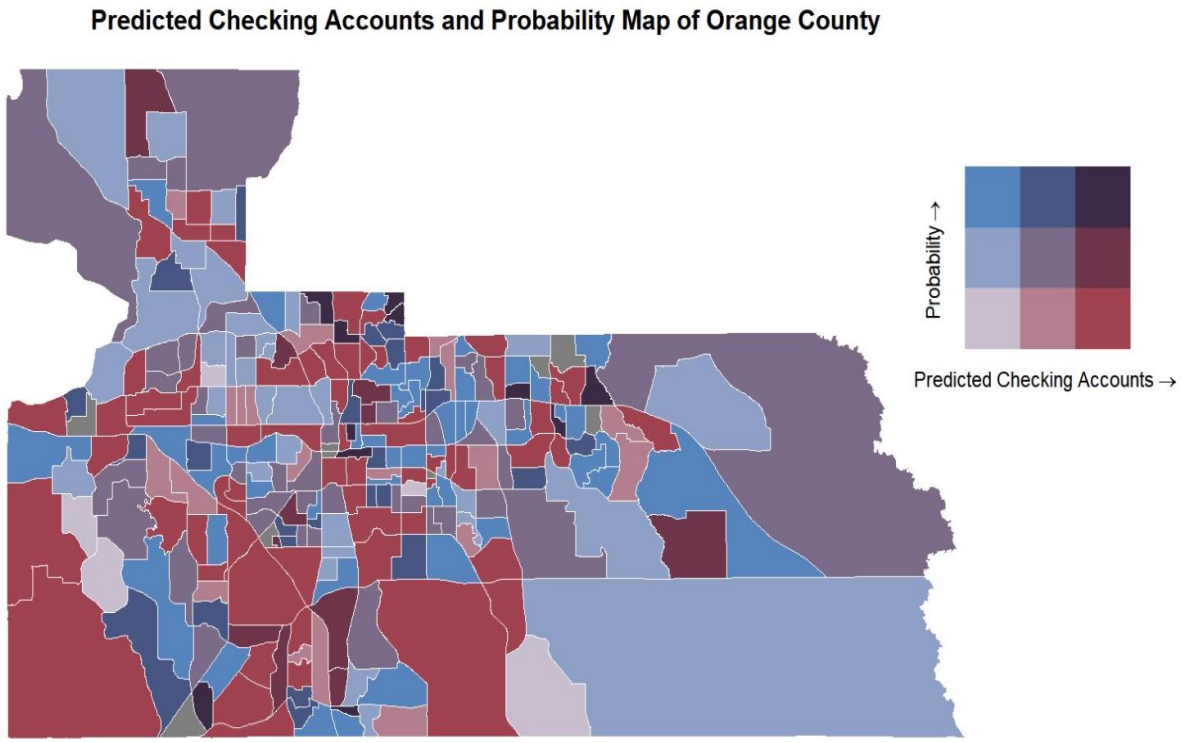


Figure 12: Predicted checking accounts and probability of being underserved for Orange County

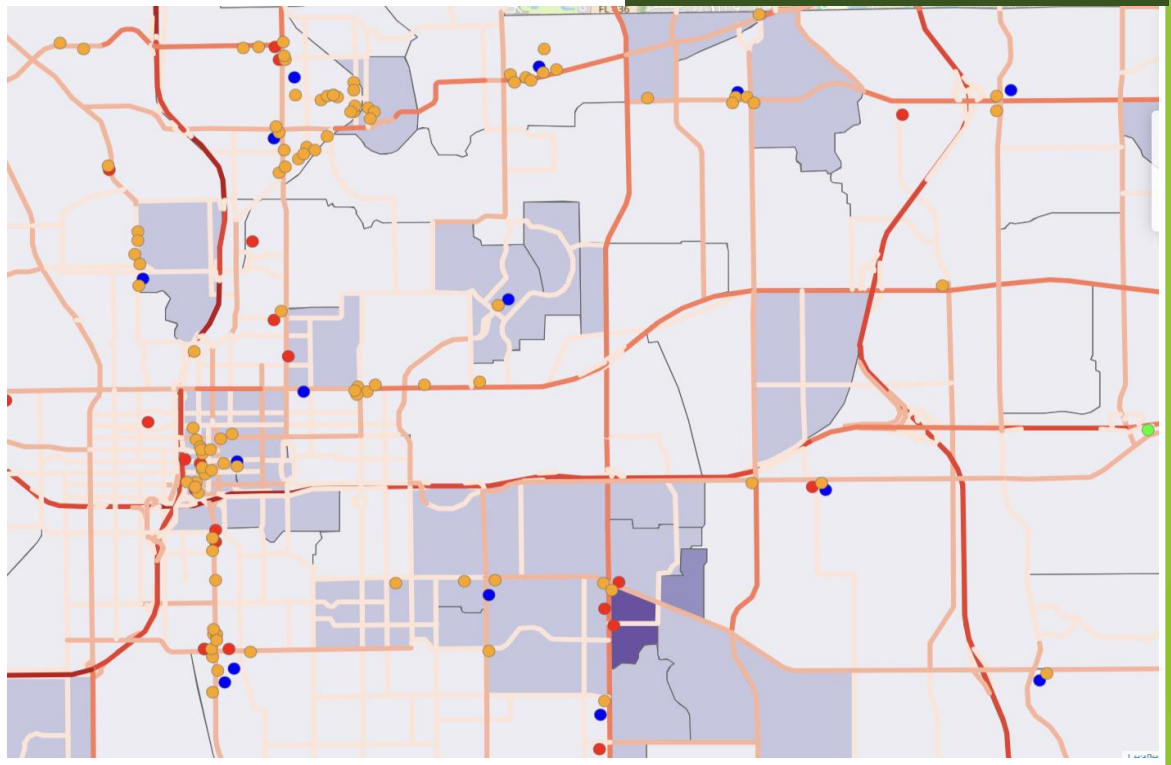


Figure 13: Area of selected location, showing population density, Publix locations, and competitor locations

Customer Needs: Achievements

- Create models that can be cross-referenced to identify potential branch sites for MidFlorida – **YES, market saturation model, Heckman**
- Identification of variables that affect profitability and the success of a branch – **YES, population, median home value, etc.**
- The model is simple enough to explain to higher management – **YES, visuals and heat maps created for high interpretability**
- Finding data that is useful and available for determining profitability – **YES, web scraping added as data gathering tool**
- Balancing a model that can be as accurate as possible while not overfitting – **PARTIAL, using multiple approaches ensures that we do not rely on one method, but we have no true "test" set to see how well our chosen location actually performs**

Further Recommendations

- Adding a temporal component would greatly improve predictive power of model
  - How the current locations have changed over time and what changes in predictors have caused changes in profitability
- Identifying better instrumental variables could make Heckman Correction possible as a modeling technique
- Predict the market saturation in next 3 years (growth in Florida and changes in demographics)

Thank You!
Questions?

25